Microbiology (2006), **152**, 1951–1968 DOI 10.1099/mic.0.28861-0

Two novel conjugative plasmids from a single strain of *Sulfolobus*

Gaël Erauso,[1,2] Kenneth M. Stedman,[3,4] Harmen J. G. van de Werken,[1] Wolfram Zillig[3]† and John van der Oost[1]

[1]Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands

[2]UMR CNRS 6539, IUEM, Université de Bretagne Occidentale, Technopôle Brest-Iroise, Place Copernic, 29280 Plouzané, France

[3]Max-Planck-Institut für Biochemie, Martinsried, Germany

[4]Biology Department, Portland State University, Portland, OR 97207, USA

Correspondence
Gaël Erauso
gael.erauso@univ-brest.fr

Received 22 January 2006
Revised 4 March 2006
Accepted 6 March 2006

Two conjugative plasmids (CPs) were isolated and characterized from the same '*Sulfolobus islandicus*' strain, SOG2/4. The plasmids were separated from each other and transferred into *Sulfolobus solfataricus*. One has a high copy number and is not stable (pSOG1) whereas the other has a low copy number and is stably maintained (pSOG2). Plasmid pSOG2 is the first *Sulfolobus* CP found to have these characteristics. The genomes of both pSOG plasmids have been sequenced and were compared to each other and the available *Sulfolobus* CPs. Interestingly, apart from a very well-conserved core, 70 % of the pSOG1 and pSOG2 genomes is largely different and composed of a mixture of genes that often resemble counterparts in previously described *Sulfolobus* CPs. However, about 20 % of the predicted genes do not have known homologues, not even in other CPs. Unlike pSOG1, pSOG2 does not contain a gene for the highly conserved PlrA protein nor for obvious homologues of partitioning proteins. Unlike pNOB8 and pKEF9, both pSOG plasmids lack the so-called clustered regularly interspaced short palindrome repeats (CRISPRs). The sites of recombination between the two genomes can be explained by the presence of recombination motifs previously identified in other *Sulfolobus* CPs. Like other *Sulfolobus* CPs, the pSOG plasmids possess a gene encoding an integrase of the tyrosine recombinase family. This integrase probably mediates plasmid site-specific integration into the host chromosome at the highly conserved $tRNA^{Glu}$ loci.

INTRODUCTION

Sulfolobus solfataricus was one of the first organisms to be recognized as a member of the Archaea (Zillig *et al.*, 1980). Due to this early identification, *S. solfataricus* and its relatives have become model organisms for fundamental studies of Archaea. Studies of the genus *Sulfolobus* have been instrumental in understanding archaeal mechanisms of transposition (Martusewitsch *et al.*, 2000), transfection (Schleper *et al.*, 1992), transformation (Aravalli & Garrett, 1997; Cannio *et al.*, 1998; Elferink *et al.*, 1996; Stedman *et al.*, 1999) and conjugation (Reilly & Grogan, 2001; Schleper *et al.*, 1995). An impressive variety of mobile genetic elements has recently been discovered in Archaea in general, and in *Sulfolobus* in particular: viruses, autonomous insertion sequence (IS) elements, non-autonomous miniature inverted repeat transposable elements (MITEs), small non-conjugative plasmids and large conjugative plasmids (Brugger *et al.*, 2002; Prangishvili *et al.*, 2001; Rice *et al.*, 2001; Zillig *et al.*, 1998). Although there have been impressive recent developments in *Sulfolobus* genetics, this remains a bottleneck (Albers *et al.*, 2006; Bartolucci *et al.*, 2003; Jonuscheit *et al.*, 2003; Stedman *et al.*, 1999; Worthington *et al.*, 2003).

The first archaeal conjugative plasmid (CP), pNOB8, was isolated from a Japanese *Sulfolobus* isolate (Schleper *et al.*, 1995). Since then, several other CPs have been isolated from colony-cloned strains of '*Sulfolobus islandicus*', and subsequently characterized (Greve *et al.*, 2004; Stedman *et al.*, 2000). Sequence comparison of all *Sulfolobus* CPs revealed three distinct sequence domains. One well-conserved cluster

†Deceased.

Abbreviation: CP, conjugative plasmid.

The GenBank/EMBL/DDBJ accession numbers for the sequences of the pSOG plasmids are DQ335583 (pSOG1) and DQ335584 (pSOG2).

An alignment of the *Sulfolobus* CP integrases with representative members of the tyrosine recombinases is available as supplementary data with the online version of this paper.

of genes covering approximately 12 kbp of the plasmids' genomes apparently contains the conjugative functions. A second is the putative origin of replication. Finally there is a region proposed to encode replication proteins (Greve *et al.*, 2004). Only a few distant homologues to bacterial proteins involved in conjugative transfer (TraG, TrbE) and partitioning (ParA, ParB) have been found. In the case of the pNOB8 and pING plasmids, derived variant plasmids were detected upon propagation. These occur as a result of deletion and recombination (She *et al.*, 1998; Stedman *et al.*, 2000). Comparing the conserved sequences of CPs with some non-conjugative derivatives has provided insight into proteins and DNA sequence motifs putatively involved in conjugation in Archaea.

A single strain of '*S. islandicus*' SOG2/4 was found to harbour two very different but related plasmids. One of these had a stable low copy number in the well-characterized *S. solfataricus* P1 strain, so it was of interest for the development of genetic tools. The two plasmids were separated and characterized. Here we present the complete sequences of these two archaeal CPs (pSOG1 and pSOG2). Comparison of these novel CPs with the available counterparts has been used to further identify plasmid features that play key roles in conjugative transfer in Archaea.

METHODS

***Sulfolobus* growth, DNA isolation and analysis.** Single-colony isolates of strains containing the pSOG plasmids were obtained and grown in standard *Sulfolobus* medium as described previously (Zillig *et al.*, 1994). Plasmid DNAs were prepared from 4 litres of freshly conjugated cells (1 : 10 000 donor to recipient followed by growth for 48 h) by using a variation of the alkaline lysis method of Birboim & Doly (1979) as described previously (Arnold *et al.*, 1999). Total DNA, i.e. chromosomal plus plasmid DNA, was isolated as described by Arnold *et al.* (1999). For electrophoretic analysis, about 3 μg total DNA or 1 μg plasmid DNA was digested with the appropriate restriction enzyme and separated on 0·6–1·0 % agarose gels (Sambrook *et al.*, 1989). Southern hybridizations were performed by using the DIG labelling and detection kit from Roche Diagnostics, according to the manufacturer's instructions. The copy number of the pSOG plasmids was determined by estimation of the ratio of single-copy chromosomal fragments to plasmid fragments in restriction digests of total DNA as described previously (Schleper *et al.*, 1995).

Cloning and sequencing. Prior to cloning, plasmid DNA preparations were purified by ultracentrifugation in a caesium chloride gradient in the presence of ethidium bromide (1 mg ml^{-1}) (Sambrook *et al.*, 1989). Digestion of both plasmids with *Eco*RI produced 11 bands for pSOG1 and 10 bands for pSOG2, ranging from 0·3 to 7·2 kbp. All of these fragments were cloned in the *Eco*RI site of pUC28 (Benes *et al.*, 1993). Fragments obtained by digestion with *Bam*HI, *Hin*dIII, *Pst*I and *Xba*I in the size range 0·8–4·5 kb were also cloned in the corresponding sites of pUC28 to obtain an overlapping clone library for pSOG1 and pSOG2. Sequencing reactions were carried out on a LiCor DNA sequencer 4000L with a Thermo Sequenase fluorescent-labelled primer cycle sequencing kit (Amersham Biosciences) and infrared-labelled primers M13 forward and M13 reverse (MWG-Biotech). Gaps in the sequence were filled by using specific primers either directly for sequencing on library clones or to sequence PCR amplicons obtained with native pSOG DNA as template. The sequences were trimmed and assembled using the SeqMan II program (Lasergene package), with both strands completely sequenced and with a minimum threefold coverage.

Computer analysis. DNA sequences were analysed using Vector NTI software (version 9, Informax). Direct and inverted sequence repeats were detected by using the GeneQuest program (Lasergene). Cumulative GC skews were made with the Genskew software (http://mips.gsf.de/services/analysis/genskew) and the Z-curve program (http://tubic.tju.edu.cn/zcurve/). Analyses were done with a window size of 30 nt. Identification of putative genes and operons was performed using the FGENESB pattern/Markov chain-based prediction program from Softberry (http://softberry.com/berry.phtml) and the pre-trained parameters of *Sulfolobus solfataricus* and *S. tokodaii*. Putative promoters (TATA box), Shine–Dalgarno sequences and terminators were identified with a window size of 12, 6 and 11 nt, respectively, in the 50 nt sequences upstream or downstream of the predicted gene start and stop codon. The nucleotide sequences were analysed using the Gibbs sampler algorithm (Thompson *et al.*, 2003). Sequence logos were generated using WebLogo (Crooks *et al.*, 2004). Homology searches were performed with a range of BLAST tools at the NCBI server (http://www.ncbi.nlm.nih.gov/blast). Identities were calculated with the program LALIGN at the Swiss EMBnet node server (http://www.ch.embnet.org/. Combined searches of a number of databases of protein families, domains and functional sites were performed using SMART (http://smart.embl-heidelberg.de/) and CDD tool (NCBI). The program COILS (EMBnet) was used for finding α-helical coiled-coil domains. Transmembrane domains were predicted by the programs PSORT (http://psort.nibb.ac.jp/), TMPRED (EMBnet) and TMHMM (http://www.cbs.dtu.dk/services/TMHMM). Identification of potential signal peptides was done with SIGNALP (http://www.cbs.dtu.dk/services/SignalP). For phylogenetic analyses, the deduced amino-acid sequences of the largest conserved ORFs in each *Sulfolobus* CP were aligned using MUSCLE (Edgar, 2004) and revised manually. Trees were generated from each individual alignment and for concatenated alignments of several ORFs, using the neighbour-joining method (Saitou & Nei, 1987) of the MEGA 3.1 program (Kumar *et al.*, 2004). Distances were calculated using the Poisson correction (PC) distance model (Nei & Kumar, 2000). Tree significance was assessed by bootstrapping 1000 times.

RESULTS AND DISCUSSION

Origin of the pSOG2/4 plasmids

Strain SOG2/4 harbouring the conjugative plasmid (CP) pSOG2/4 was isolated from samples collected in the Sogasel Icelandic solfataric field and belongs to a species provisionally called '*Sulfolobus islandicus*', closely related to *S. solfataricus* (Zillig *et al.*, 1994). The conjugative nature of this plasmid was shown by its capacity to be directly transferred from donor into recipient cells, resulting in complete spread through the recipient culture (Prangishvili *et al.*, 1998). Upon conjugation into a foreign host the plasmid was amplified to high copy number (more than 35 copies per chromosome), as observed for the other *Sulfolobus* CPs. Unlike other *Sulfolobus* CPs, which were mostly stable immediately after conjugation, the plasmid from SOG2/4 appeared to be very unstable when transferred into *S. solfataricus* strain P1 but remained indistinguishable from the original when another '*S. islandicus*' strain, HVE10/4, was the recipient (Prangishvili *et al.*, 1998) (Fig. 1). However in the former case, as often observed for other

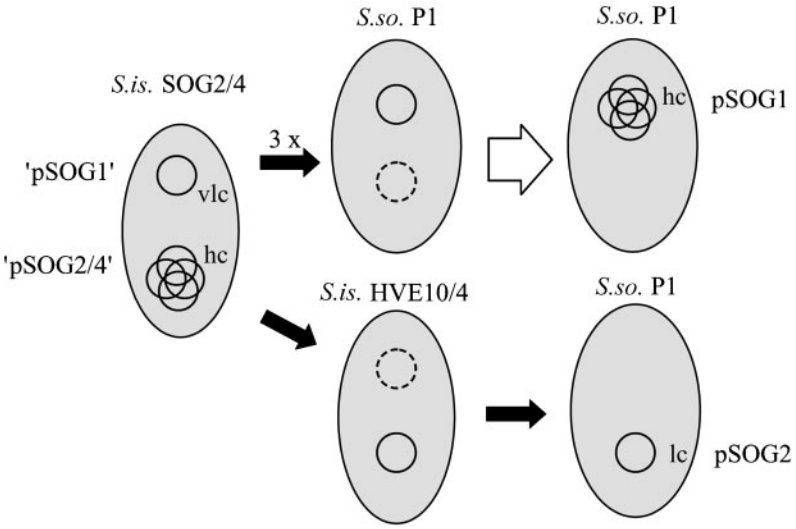

Fig. 1. Generation of plasmids pSOG1 and pSOG2. Plasmid hosts are shown as large ovals. Plasmids are shown as circles; a dashed-lined circle indicates plasmid loss. Black arrows represent plasmid transfer by conjugation. The broad white arrow represents single colony isolation. 'lc' indicates low copy number; 'hc' indicates high copy number; 'vlc' indicates very low copy number, which may be an integrated copy. 'pSOG1' is the precursor of the plasmid pSOG1, which contains putative conjugation genes and origin of the pKEF family of *Sulfolobus* CPs. 'pSOG2/4' is the originally observed plasmid and precursor of plasmid pSOG2.

Sulfolobus CPs, prolonged growth of transcipients resulted in plasmid variant formation and eventual curing (Schleper *et al.*, 1995; She *et al.*, 1998). Plasmid pSOG1 (previously named pSOG2/4 clone 1; Prangishvili *et al.*, 1998) was isolated from a single colony obtained by plating transcipient cultures from a third successive conjugative transfer in *S. solfataricus* strain P1 (Prangishvili *et al.*, 1998). A stable clone, named pSOG2 (previously named pSOG2/4 clone A), was obtained by conjugative transfer first in strain HVE10/4 and subsequentely in *S. solfataricus* P1; pSOG2 appeared to be indistinguishable by restriction endonuclease digestion from the original pSOG2/4 CP. See Fig. 1 for an overview of the isolation procedure. Comparison of the restriction endonuclease digestion patterns and Southern blotting of pSOG1 and pSOG2 showed that the two plasmids differ dramatically (not shown; and see Prangishvili *et al.*, 1998). Only 1/3 of the sequence of pSOG2 is conserved in pSOG1 (Fig. 2). The rest of the pSOG1 genome was likely acquired via recombination from another low-copy-number plasmid in the parent strain SOG2/4 that was lost upon passage through the '*S. islandicus*' HVE10/4 strain. The 'pSOG1' plasmid must have been present in the original strain at very low copy number because it was not detected by Southern blotting that detected low-copy-number plasmids. After long exposures, background hybridization to genomic DNA was detected that may indicate an integrated plasmid (not shown). After passage, pSOG2 can be stably propagated in *S. solfataricus* P1. Copy number control appeared to be lost or damaged in the pSOG1 variant, as indicated by its extremely high copy number in *S. solfataricus* P1 as compared to the low copy number of pSOG2.

Nucleotide sequence of the pSOG1 and pSOG2 plasmids

The assembled circular sequences are 29 000 bp in length for pSOG1 and 26 960 bp for pSOG2 (Fig. 2). Their overall G+C contents were 35·8 mol% and 36·7 mol% respectively. The corresponding value for the chromosome of '*S. islandicus*' is not known but the value determined for *S. solfataricus* (35·8 %) (She *et al.*, 2001) is identical to that of pSOG1. As previously deduced from their *Eco*RI restriction patterns and Southern hybridizations, the two genomes share a large 100 % identical region of 9842 bp (nucleotides 14 696–21 466 and 24 217–27 129 for pSOG1, 15 815–25 657 for pSOG2). This region in pSOG1 is interrupted by a non-homologous sequence of 2756 bp. As expected from previous studies (Greve *et al.*, 2004; Stedman *et al.*, 2000), pSOG2 shares extensive nucleotide sequence similarity (long stretches of sequences up to ~3 kbp with more than 95 % identity) with other *Sulfolobus* CPs of the pKEF group (nomenclature according to Greve *et al.*, 2004), whereas pSOG1 has more similarities with the pARN group of plasmids, which also contains plasmids integrated into the genomes of *Sulfolobus acidocaldarius* and *S. tokodaii* (Chen *et al.*, 2005; She *et al.*, 2004).

The G+C content of pSOG CPs is not evenly distributed, displaying a number of peaks and troughs (not shown). Five regions of more than 2000 bp have a higher G+C content (>36 mol%); these fragments roughly correspond to parts of the genome that encode the most-conserved ORFs in *Sulfolobus* CPs (Fig. 2, Table 1). In contrast, lower G+C regions are less extended and contain less-conserved ORFs. The latter fragments may encode functional units, such as partitioning and additional elements involved in conjugation (see below), also indicating that pSOG2/4 plasmids have a mosaic structure composed of elements of diverse origin. A clear minimum, corresponding to several successive short poly(A) stretches, is located just in front of ORF175, present in both plasmids (ORFs present in only one plasmid are listed as ORF1- for pSOG1 and ORF2- for pSOG2).

ORF distribution

Forty-six ORFs encoding a product at least 50 aa in length were identified in the genome of pSOG1 and 41 ORFs

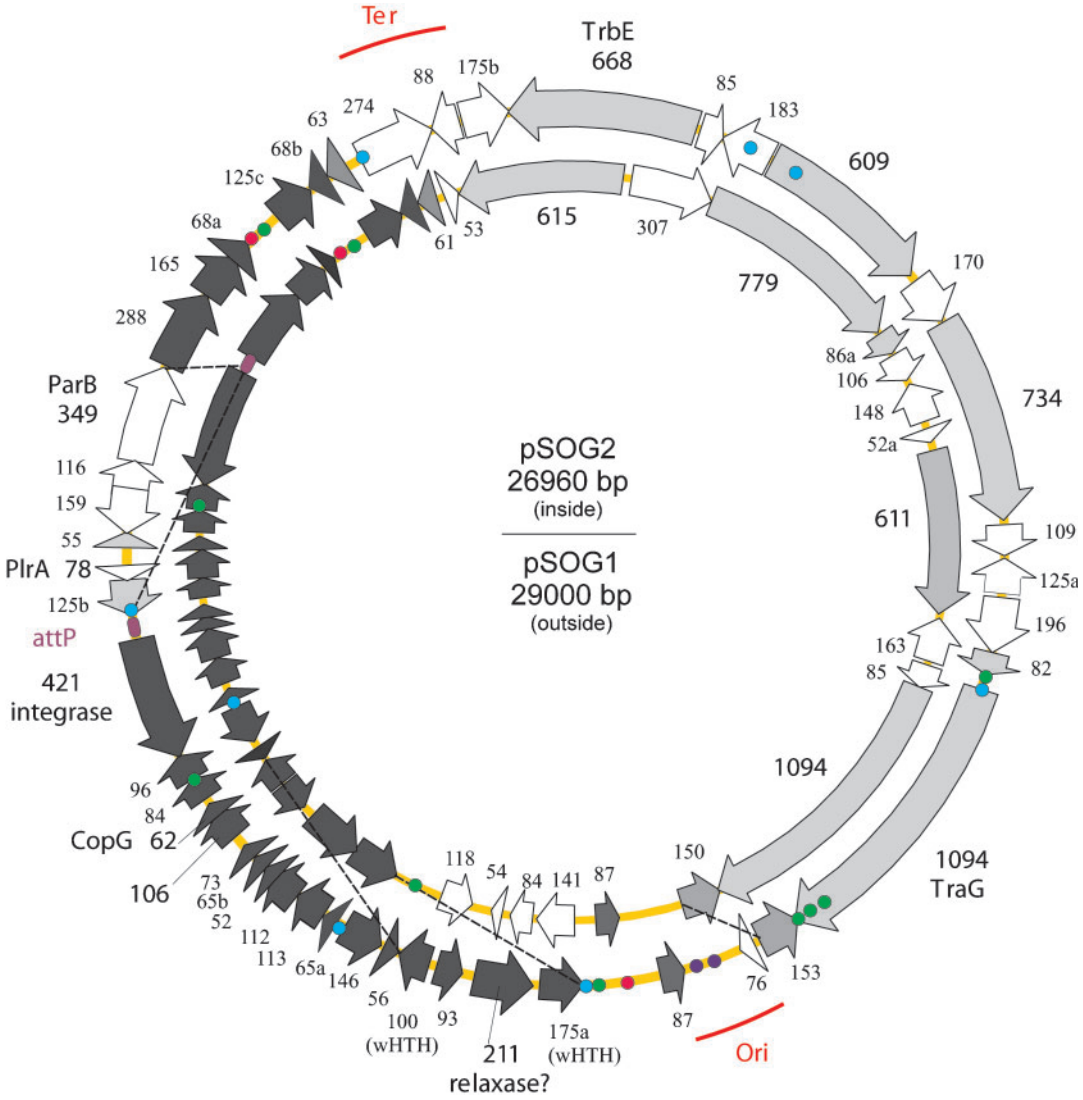

Fig. 2. Comparison of the pSOG1 and pSOG2 sequences. This diagram shows the circular genomes of pSOG1 on the outside and pSOG2 on the inside. ORFs are shown as arrows. Similar ORFs in the two plasmids are filled in grey; identical ORFs are filled in black; ORFs not conserved between the two plasmids are not filled. ORFs with predicted functions are labelled and ORFs discussed in the text are in bold. Insertions and gene replacements are indicated by dashed lines between the two genomes. ORF names are shown next to the corresponding arrows. The recombination motif TAAACTGGGGAGTTTA is represented by a small disk, coloured green when present on the direct DNA strand and light blue when located on the complementary strand. Blue disks indicate the two larger tandem repeats, and a red disk indicates larger inverted repeats. The violet oval represents the putative site of integration *attP*. The approximate location of the origin (Ori) and terminus (Ter) of replication as predicted by cumulative GC skew and Z-curve analyses are also indicated.

in the genome of pSOG2 (Fig. 2, Table 1). These putative genes are generally closely spaced (mean density 1·55 ORFs per kb). However, two regions with larger intergenic spaces are located at the borders between the conserved and the variable regions of the two pSOG plasmids: between ORFs 1-76 and ORF87 and from ORF87 to ORF175a on one side, and between 1-68a and 1-125c on the other side. The putative genes are almost equally distributed on both DNA strands and apart from a few clusters of genes that are probably co-transcribed (see below) the genes are evenly dispersed. This ORF distribution is also observed for the other *Sulfolobus* CPs (Greve *et al.*, 2004) (see below). Another common feature of the plasmids is that all of the larger ORFs (encoding >500 aa) are located in the same conserved region of their genome (Fig. 2, Fig. 4).

Operons and putative transcriptional and translational signals

The pSOG plasmid ORFs start at ATG (79·5 %), TTG (17·5 %) or GTG (3 %) and terminate at TAA (44 %), TGA (39 %) or TAG (17 %). This distribution of start and stop codons resembles that of the *S. solfataricus* chromosome (Garcia-Vallve *et al.*, 2003). Except for the above-mentioned two regions containing larger intergenic sequences, in 83 % of the cases an ORF is found within 50 nt of the previous ORF's stop codon. Moreover, for 31 % of the collinear ORFs this distance is less than 20 nt (75 % of these overlap), and the latter have been considered to be part of an operon. Sequence logos derived from alignment of the 50 nt upstream sequences of pSOG genes allowed us to identify putative translational and transcriptional signals (Fig. 3). The consensus ribosome-binding site (GGTGA) was found in all but a few genes that are assumed to be part of an operon (Table 1, Fig. 3). It is optimally located at positions −10 to −7 bp upstream of the putative start codon. This sequence is the reverse complement of (underlined) part of the 3′ end of the 16S rRNA sequence from *S. solfataricus* (GGAUCACCUCA-3′). However, such a sequence was not detected for single genes or first genes of a candidate operon, confirming the results of previous analyses done on *Sulfolobus* (Tolstrup *et al.*, 2000) and later on a large set of archaeal genomes (Torarinsson *et al.*, 2005). Accordingly, for this class of genes, we also found a 7–8 nt A+T-rich sequence centred between positions −25 and −27 from the start codon, fulfilling the criteria for the *Sulfolobus* TATA box of Soppa (1999). The promoter sequences of *S. solfataricus* generally contain a transcription factor B responsive element (BRE) with two to four A(T)s generally located 2 nt upstream of the TATA box sequence (Bell & Jackson, 2000); such a conserved BRE could generally be identified in both pSOG plasmids (Table 1, Fig. 3). The distance between the predicted TATA boxes and the putative start codon coincides with the mean interval found experimentally between the TATA box and the transcriptional start in mapped *Sulfolobus* promoters (Dalgaard & Garrett, 1993; Reiter *et al.*, 1988). This means that there is little or no room for a ribosome-binding site and explains why this signal was not found in our analysis of single genes and first genes of an operon. It also implies that translation initiation for this class of genes must depend on a mechanism other than Shine–Dalgarno sequence (Condo *et al.*, 1999). Table 1 summarizes the information obtained from transcriptional and translational signal searches: pSOG plasmids appear to be organized in 30 transcription units (TU) for pSOG1 (5 operons and 25 single genes) and 26 TU (6 operons and 20 single genes) for pSOG2. Additional support for the co-transcription of the proposed TUs is provided by the identification, downstream of the last gene of the TU, of potential transcriptional terminators identical to those found in the virus SSV1 (Palm *et al.*, 1991) and in the *Sulfolobus* chromosome (She *et al.*, 2001) e.g. 5′-TTTTTT or 5′-TTTTCTT or 5′-TTTATTTT. The fraction of single genes (69 %) is quite high compared to that found in other *Sulfolobus* extrachromosomal elements (e.g. 27 % for the genome of virus SSV1; Palm *et al.*, 1991). This mosaic character may reflect the need for fine-turning each gene expression separately and/or the modularity of these plasmids.

Overall genome comparison with the other Sulfolobus CPs

Similarity searches showed that 53 of the 65 unique ORFs (80 %) of pSOG1 and pSOG2 had significant matches (BLASTP E-value $<10^{-4}$) to proteins in public databases. Most of the hits were to hypothetical proteins encoded by other *Sulfolobus* CPs, showing from 26 % and up to 100 % amino acid sequence identity (Table 1). Ten homologous ORFs are shared by the eight CPs, while over 80 % of the other ORFs are common to two or more CPs. As illustrated in Fig. 4, the conserved ORFs are clustered in two genomic regions separated by a larger intergenic section. These three genomic sections, named A, B and C according to Greve *et al.* (2004), appear to be functionally distinct. The largest one, section A, also contains the highly conserved large ORFs: 1-668 (TrbE), 1-609, 1-734 and 1-1063 (TraG) (pSOG1 numbering) which are most likely involved in conjugation. Section B carries the putative origin of replication. Section C corresponds to a cluster of closely packed genes, including the six other genes common to all CPs: a putative relaxase (211), an operon containing genes implicated in plasmid replication, 106 (RepA), 62 (CopG), 421(integrase), 84 and 93b (two hypothetical proteins). The sequence of an apparently defective CP, pTC, from *Sulfolobus tengchongensis* (Xiang *et al.*, 2003) has been deposited in GenBank (NC_005969). It is missing a number of conserved ORFs from the other CPs, specifically a homologue of the highly conserved pSOG ORF2-779 (which appeared to be partitioned in three ORFs) and putative replication ORFs including all of sections B and C (Fig. 4). We are looking forward to publication of details on its isolation and physical characteristics.

To better evaluate the relationship between the eight self-transmissible plasmids of *Sulfolobus*, we used the most representative genes as phylogenetic markers. Unrooted trees were obtained from alignment of the eight homologous genes for each of the four largest ORFs of section A, namely TrbE, 2-779, 2-610 and TraG, and for a concatenated alignment of these four genes. The topology obtained for each individual gene (not shown) was very similar to that obtained for the concatenated tree (Fig. 5a) and not dependent on the method used for tree construction (neighbour joining, minimal evolution, parsimony). The concatenated tree clearly shows two distinct groups: (i) the pKEF9 group (Greve *et al.*, 2004), which includes pSOG2 and the more divergent pNOB8 branch, and (ii) the pARN group, to which pSOG1 may belong. However, this clustering does not agree with the integrase tree, where pARN, pHVE and pKEF appear closely related whilst the pING integrase is the most divergent. This difference in tree topology suggests a distinct origin for the respective

Table 1. Properties of ORFs and operons of plasmids pSOG1 and pSOG2

ORFs in pSOG1 and/or pSOG2				Closest relatives		Family, Domains, Motifs¶	Predicted function
Name†	Promoter motif‡	SD motif‡	Terminator motif‡	ORF-Plasmid§	Identity (%)‖		
1-668 (–)	ACCCTG**TTTATAA*** (–17)	–	(+1) TTATTTT	624-pARN3	27	COG3451: VirB4, TrbE; COG0433: ATPases; pfam01935: unknown TM (1), SP? mbc	Conjugation (mpf)
1-85	GATTTA**TTTATAA** (–23)	–	(+26) TTTTTTCTT		–		
1-183 (–)	ACCAA**TTTATTTTA** (–23)	–	(+4) CCTTTTTTCT	187-pARN3	33	TM (2), SP, im	
1-609	AATTAA**AATAAA** (–39)	–		660-pARN3	48	TM (1), SP, mbp	
1-170		tAGGTat (–10)	(+6) TTTTTC				
1-734	GGAGG**TTTAAT**TTA (–25)	–	(+9) TTTTTATTT	458-pTC507	26	COG1196: Smc, ATPases.	Conjugation (mpf)
				487-SAC3	29	Chromosome segregation	
				507-pARN3	26	TM (9), SP, im; COG0477: permeases of the major facilitator superfamily	
1-109a	AAAATAT**TTTTAT** (–25)	–	(+23) TTGTTTTT			TM (1), SP, im	
1-125a (–)	AATT**TGATTA**TTA (–21)	–	(+2) TTTTTTCTT			TM (3), SP, im	
1-196	ATAT**TTTATAA**A (–23)	–				TM (1), SP, mbp	
1-82		cAGGTGA (–9)	(+6) TTTTTTCTT	86a-pSOG2	30		
1-1063	CTCCC**ATTTTT**ATA (–24)	–	(+1) TTTTTTCCTT	1092-pHVE14	38	pfam02534: TraG/TraD family; COG3505: VirD4; COG1674: FtsK/SpoIIIE DNA segregation ATPase	Conjugation (coupling protein)
1-153 (–)	GGGTT**TTTAAA**TA (–28)	–	(+12) TTTTAATTTT	153a-pING1	93		
1-76¥	AAATA**TTTAAA**AA (–40)	–	(+5) TTTTCT	69-pARN3	75	TM (1), SP, im	
87 (–)	TTATAC**TTTATA**TA (–25)	GAGGTtA (–4)	Not found	87-pSOG2	85		
				99-pARN3	65		
175a (–)		GgGGgGt (–5)	(+31) CCTTTTTT	188-SAC3	49	TrmB family (pfam 01978) wHTH transcription regulator	Transcription regulation?
211 (–)		GgGGgGA (–3)		200-pST2 (ST2505)	64	pfam 00135: carboxyl-esterases	Mobilization?
				253-pNOB8	50		
93a (–)	AAATAGA**TTTATA** (–25)	–		109-pARN3	83	COG4742: transcriptional regulator	Transcriptional regulation
100	TATGTTC**ATTATA**T (–23)	–	(+13) TTTTTCTT	132-pST2 (ST1340)	89	COG1846.1 and pfam1047: MarR family; wHTH	Transcription regulation
				100-pHVE14	99		
56 (–)	AGTCCAC**TATTTA**T (–25)	–	(+12) TTTTT	56-pARN3	100		
146 (–)	AATTGGA**AAAAAA** (–33)	–	(+1) TTTTC	146-pARN3	97		
65a (–)	AAATA**TATAAA** (–58)	–		65-pHVE14	100		
113	AAAAGA**TTTATA** (–33)	–		102-pHVE14	90		Replication
112		GAGGTGA (–7)		79-SIFV (SIFV0017)	40		Replication
52		GAGGTGA (–6)		52-pHVE14	100		Replication

Table 1. cont.

ORFs in pSOG1 and/or pSOG2				Closest relatives		Family, Domains, Motifs¶	Predicted function
Name†	Promoter motif‡	SD motif‡	Terminator motif‡	ORF-Plasmid§	Identity (%)‖		
65b		GAGGTGg (–4)		D-63 SSV1	42		Replication
73¥		GGGcTGA (–3)		70-pHVE14	86		Replication
106		GgtGTGA (–5)		107-pHVE14	92		Replication
62		GAGGaGg (4)		62-pARN3	77	CopG family	Copy number control
84		aGGGTGc (–7)		87-pKEF9	85	pfam: DUF904, leucine-zipper	Replication
93b		GgGGTGA (–5)	Not found	109-pARN3	83		Replication
421 (–)	AAGGAT**ATTTTT** (–37)	–	(+5) TTTCCTT	419-pKEF9	63	pfam 00589 and COG4974: phage-tyrosine integrase (C-term. half); HTH-XRE: phage repressor family (N-term.)	Integration
1-125b (–)		ctGGgGA (–9)	(+43) TTTTCT	128-pARN3	90		
1-78 (–)	AACCT**ATTAAA** (–24)	–		77-pARN3 (PlrA)	98	pfam 05584: *Sulfolobus* plasmids regulatory protein	Transcription regulation
1-55¥	CAAGC**GTTAAA** (–38)	–	(+10) TTCCTTT				
1-159 (–)	GGGCA**TTTATA** (–27)	–	(+17) TTCCTT			COG1846.1: MarR family; wHTH	
1-116	AGAAT**ATATTA** (–35)	–		76-SIFV	40	TM (1), mbc	
1-349		agGGTGA (–6)		(290) SpoJ *Bacillus cereus*	35	ParB family (N-term.), DUF1130 (C-term.)	Partition
288		atGGTGA (–4)		(284) *Chloroflexus aurantiacus*	96	Conserved hypothetical	
165		caaGtGG (–5)				TM (1), im	
68a		Not found	(+33) TTTTCT	75-pKEF9	84		
125c	ATTT**TTTAAA**⋆ (–14)	–	(+1) TTTTC	126-pKEF9	82		
68b (–)	ATGATT**ATATAA** (–27)	–	(+2) TTTTCTT	68-pKEF9	75		
1-63 (–)	AGGGT**TTTAAA**AA (–23)	–		63-pARN3	65	TM (2), SP, im	
1-274	TGATA**TTTATA**A (–23)	–	(+3) TCTATTTCT				
1-88 (–)	ATTATTT**TTTATA**T (–20)	–	(+13) TTTTTT				
1-175b	AAGAG**TTTGATA**T (–23)	–	(+1) TTGTTTTTT			TM (3), SP, im	
2-615 (–)	ATAGT**ATTTAA** (–24))	–	(+3) TTTTTCTTC	630a-pNOB8	91	COG3451: VirB4, TrbE; COG0433: ATPases; pfam01935: unknown TM (1), SP? mbc	Conjugation (mpf)
2-307	GGAGTA**TTTAAA** (–27)	–		312-pNOB8	88	TM (1), mbp; CC (3) Smc motor protein	DNA mobility Conjugation? Partition?
2-779		tAGGTGg (–6)		781-pKEF9	82	TM (1), SP, mbp	Conjugation (mpf)
2-86a		GAGGaGA (–1)	(+1) CTTTTTTC	86-pHVE14	79		
2-103	GAGAGTA**TTTAAA** (–29)	–	(+36) TTTTTTT	106-pHVE14	77	TM (2), SP, im	

Table 1. cont.

ORFs in pSOG1 and/or pSOG2				Closest relatives		Family, Domains, Motifs¶	Predicted function
Name†	Promoter motif‡	SD motif‡	Terminator motif‡	ORF-Plasmid§	Identity (%)‖		
2-148 (–)	TTCTTTG**TATATA**A (–22)	–	(+26) TTTTT	150-pHVE14	94		
2-52a	TTATA**TTTATT**A (–26)	–		52-pING1	96		
2-611	TTTTTA**TTTTAA** (–23)	–	(+22) TTCTTTTTT	604a-pNOB8	65	COG1196: Smc, ATPases chromosome segregation, TM (10), SP, im; COG0477: permeases of the major facilitator superfamily	Conjugation (mpf)
2-163 (–)	AAAAGG**TTTAAA** (–23)	–	(+24) TCTTAATTT-GATTTTTTA	165-pNOB8	84	TM (1), SP	
2-85	GGTTA**TTTAAA** (–20)	–				TM (2), SP, im	
2-1094		GcGGTGA (–7)	(+25) TTTTTCT	1084-pKEF9	81	pfam02534: TraG/TraD family; COG3505: VirD4; COG1674: FtsK/SpoIIIE DNA segregation ATPase	Conjugation (coupling protein)
2-150 (–)	AGAGGTT**TTTAAA** (–27)	–	(+55) TTTTTTAT	150-pST2	55		
				153b-pING1	34		
2-87	ATATAT**TATAAA*** (–16)		(+5) TTTGT	99-pARN3	59		
2-141	GCTAAAT**AATAAA** (–26)	–	(+14) TTTTTTT	72-pARN3	77	TM (1), SP, mbp	
2-84a[¥]	GATATTA**AATAAA** (–29)	–	Not found				
2-54	TATTTT**CTTATA**T (–27)		(+14) TTTTTTTATT			TM (1), SP, im	
2-118 (–)	AATAA**TTTAAA** (–35)		(+1) TTTgcTT	120-pST2- (ST1333)	83		
2-61	AGGTAAG**ACAATA** (–23)			61-pHVE14	75		
2-53	ACGCT**TTCAAA** (–40)		(+20) TTTTTCT				

†The name of the predicted ORFs starts with '1-' for those which are unique to pSOG1 and with '2-' for those unique to pSOG2, followed by their size of their product in amino acids; (-) indicates not found. All listed ORFs were predicted by FGENESB except those marked by the symbol [¥]. Putative TATA-like promoter motifs, Shine–Dalgarno (SD) motifs and T-rich terminator sequences are shown; nucleotides which fit to the canonical TATA motif are in bold. * Indicates a putative TATA box located too close or too far from the translation start. Putative operons are indicated by a vertical bar on the left of the corresponding ORF names.

‡Numbers in parentheses indicate the position of the last nucleotide of the promoter and SD sequences, and the first nucleotide of the terminator sequence, relative to the predicted start or stop codon, respectively.

§Homologous ORFs in other CPs from '*S. islandicus*' (pING1, pARN3, pKEF9, pHVE14), *S. japonicus* (pNOB8), '*S. tengchongensis*' (pTC), or in integrated conjugative elements of *S. tokodaii* (pST2) and *S. acidocaldarius* (SAC3).

‖Percentage of amino acid identity calculated after a pairwise global alignment produced by the program ALIGN when possible or by local alignement (PSI-BLAST).

¶HTH, helix–turn–helix motif; wHTH, winged helix–turn–helix; TM (*n*), *n* transmembrane segments; SP, signal peptide; im, integral membrane protein; mbc, membrane-bound cytoplasmic; mbp, membrane-bound periplasmic; CC, coiled-coiled domain. Databases used: COG, cluster of orthologous genes; pfam, protein family; DUF, domain of unknown function, which may concern either the N-terminal (N-term.) or the C-terminal (C-term.) part of the predicted protein.

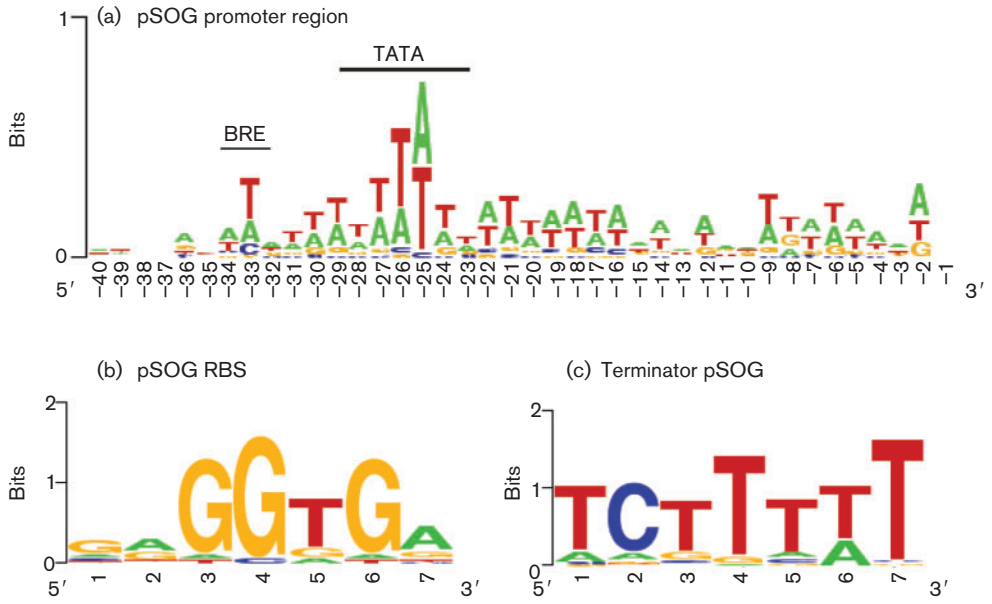

Fig. 3. Sequence logos of putative promoters, ribosome-binding sites (RBS) and terminators of pSOG plasmids. (a) Upstream pattern sequence of single genes and putative first genes of an operon (46 sites); the approximate locations of the BRE motif and TATA box are indicated by horizontal bars. (b) Putative ribosome-binding site of genes within an operon (22 sites). (c) Putative terminator pattern (46 sites).

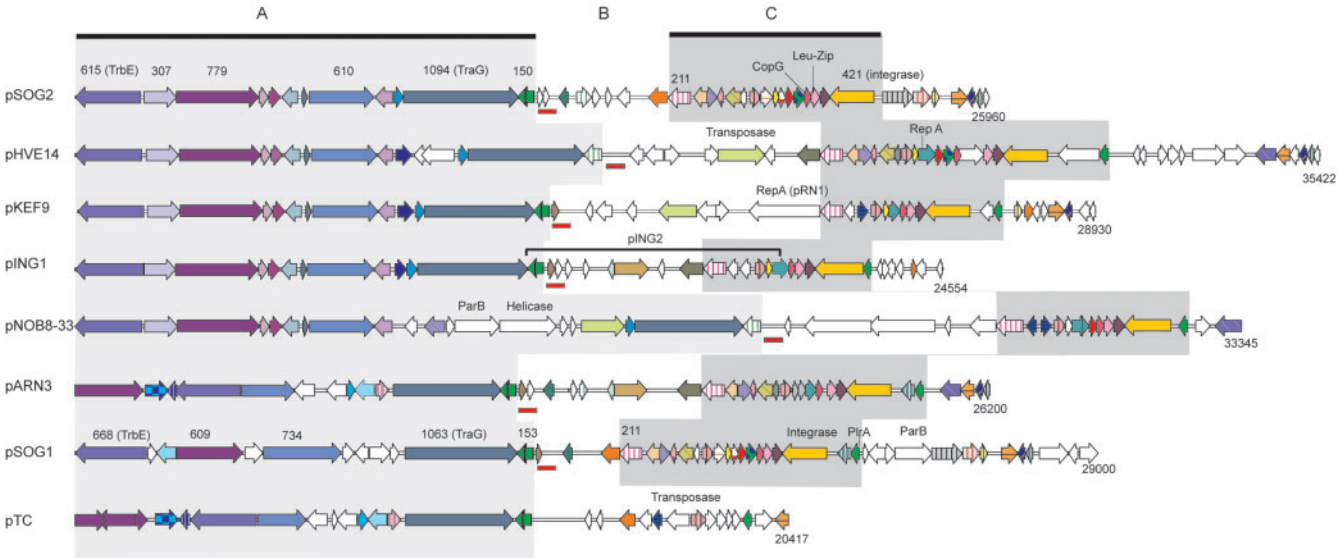

Fig. 4. Comparison between pSOG1 and pSOG2 and other *Sulfolobus* CPs. Genome maps are shown for all of the published *Sulfolobus* CPs and for the presumed defective plasmid pTC (see text for details). Homologous ORFs in different plasmids can be identified by colour and pattern. ORFs represented by white arrows have no homologues in the other *Sulfolobus* CPs. Predicted regions encoding conjugative functions are shown as region A, the putative replication origin in region B is delimited by a short red horizontal bar and the putative replication region is labelled C (figure modified from Greve *et al.*, 2004 with permission). ORFs discussed in the text are labelled.

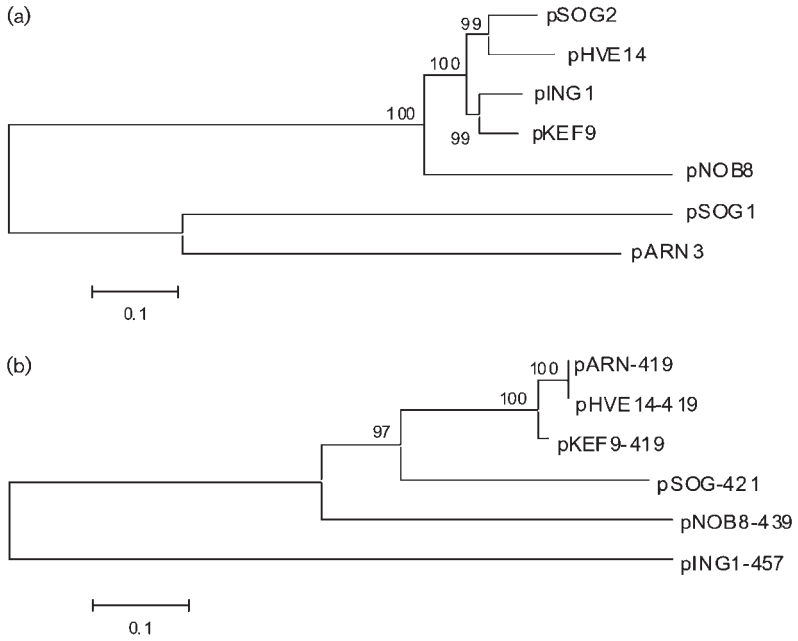

Fig. 5. Phylogenetic relationships between large conserved ORFs of *Sulfolobus* CPs. (a) Unrooted tree obtained from an alignment of the concatenated four most conserved ORFs of *Sulfolobus* CPs, presumably involved in conjugation: TrbE, TraG, ORF734 and ORF609 (pSOG1 numbering). (b) Unrooted tree of the integrase gene of *Sulfolobus* CPs. Trees were constructed using the neighbour-joining method in MEGA 3.1 (Kumar *et al.*, 2004). Branches are labelled with their corresponding bootstrap values (only those greater than 50 % are indicated).

genomic fragments coinciding with their functional modularity, section A being devoted to conjugation and section C to replication and recombination.

Conjugative transfer function

At least three of the largest ORFs in the pSOG CPs are probably involved in the conjugation process. As previously reported for their homologues in other *Sulfolobus* CPs (Greve *et al.*, 2004; She *et al.*, 1998; Stedman *et al.*, 2000), ORFs 1-1023/2-1082 showed significant similarities with the TraG/VirD4 [cluster of orthologous groups (COG) 3505], and ORFs 1-668/2-615 with TrbE/VirB4 (COG 0433). Both TraG and TrbE represent families of ATPases that are involved in conjugation in bacteria (http://www.ncbi.nlm.nih.gov/COG/) (Grohmann *et al.*, 2003). These two proteins aligned with each other around the type I ATP-binding site (Walker A motif), which occurs at a similar position in each protein (not shown). The TraG and TrbE proteins have been proposed to be coupling proteins (Grohmann *et al.*, 2003) connecting the relaxosome, a DNA-binding protein-complex encoded by both the CP and the host chromosome at the plasmid transfer origin *oriT* (Lanka & Wilkins, 1995), to the mating-pair formation (mpf) system, a plasmid-encoded multi-protein complex that is involved in the transfer of the donor DNA to the recipient cell (Llosa *et al.*, 2003). In the current model, TraG is a membrane-anchored, multimeric protein forming a pore-like structure that actively exports the transferred DNA (T-DNA) via envelope-spanning mpf components (Llosa *et al.*, 2002, 2003; Llosa & de la Cruz, 2005; Schroder *et al.*, 2002). Accordingly, we found that both TraG-like ORFs 1023/1082 and TrbE-like ORFs 668/615 possess a predicted N-terminal transmembrane domain that could serve as an anchor. They also have the same predicted topology as their bacterial homologues (Schroder & Lanka, 2003). A third large ORF in the pSOG plasmids is also highly conserved in other *Sulfolobus* CPs (Table 1). These ORFs 1-734 and 2-610 share significant similarities with permeases of the major facilitator superfamily (PSI BLAST with E-value 10^{-47} after 5th iteration). Its product possesses up to 12 putative transmembrane segments covering two-thirds of the protein sequence. It also contains a type I ATP-binding site located roughly in the same region as the TraG-like and TrbE-like ORFs. This motif is part of a conserved domain (COG1196) typical of motor ATPases, including the Smc proteins, involved in chromosome segregation or compaction (Elie *et al.*, 1997).

By analogy with the model recently proposed for bacteria (Grahn *et al.*, 2000), we assume that in *Sulfolobus* conjugation, the TraG-like and the TrbE-like proteins form a heteromultimeric complex associated with the cytoplasmic membrane, and pump the DNA through a membrane-spanning channel constituted of at least the permease-like component, which may contribute actively to the T-DNA translocation. Several other pSOG ORFs, including the fourth largest, 1-609 and 2-615, contain putative membrane helices with predicted inner and outer segments (Table 1), and may also be involved in mating pair formation (mpf). In bacterial CPs, genes encoding conjugative transfer functions are generally clustered in one or two *tra* regions (Grohmann *et al.*, 2003; Pansegrau *et al.*, 1994), one encoding the relaxosome and the other the mpf system. The remarkable conservation of the almost contiguous cluster of ORFs (including the putative TraG, TrbE, ORFs 1-734/2-610 and 1-609/2-615) in all the *Sulfolobus* CPs suggests that these genes are involved in the same function, probably the mpf system.

Homology searches failed to detect putative components of the relaxosome in the pSOG sequences. Previous comparison of the genome sequence of functionally defective pING variants that had lost their capacity for self-transfer but were still transmissible led to the proposal that some of the conserved ORFs encode mobilization (*mob*) functions (Stedman *et al.*, 2000). Among the four candidate *mob* genes only one, ORF211, is present in all self-transmissible plasmids of *Sulfolobus* CPs (Fig. 4). It is therefore tempting to speculate that, as in bacteria (Francia *et al.*, 2004), archaeal mobile DNA elements carry the information necessary for relaxosome formation. The small pING plasmids of *Sulfolobus* should contain a gene encoding a functional analogue of the bacterial relaxases as well as a *cis*-acting transfer origin (*oriT*). Previous attempts to locate *oriT* in *Sulfolobus* CPs showed that six conserved sequence motifs could potentially play that role (Stedman *et al.*, 2000). We found that only one of these sequence elements ('motif 2') is conserved in the pSOG plasmids and generally in the *Sulfolobus* CPs of the pKEF9 family (but not in the pARN plasmids), as well as in mobilizable pING derivatives, and that this motif is always located immediately upstream of ORF211. In a typical bacterial mobilization region, *oriT* is located upstream of the gene encoding the relaxase (Francia *et al.*, 2004). The genomic context suggests a hypothetical function of relaxase for ORF211. However, the lack of any detectable sequence similarities with bacterial Mob proteins makes this assumption questionable.

Partitioning and plasmid maintenance

The N-terminal half (amino acids 1–160) of ORF1-349 in pSOG1 is similar to the highly conserved ParB/SpoB protein family involved in partitioning of bacterial plasmids and chromosomes (Table 1). The sequence includes two conserved motifs that are proposed to be involved in interaction with ParA/SopA and unknown host factors in bacteria (Hanai *et al.*, 1996), and a helix–turn–helix DNA-binding domain typical of the ParB family. The three motifs aligned well with a set of divergent bacterial ParB proteins but poorly with ORF470 and ORF422 of the *Sulfolobus* plasmid pNOB8 (She *et al.*, 1998), and not at all to ORFs in other *Sulfolobus* CPs (not shown). Due to (i) the relatively high similarity of ORF1-349 to bacterial rather than archaeal homologues, (ii) the significant difference in codon usage compared to the other pSOG ORFs (data not shown), and (iii) the genomic context (Fig. 2), it is suspected that ORF1-349 and surrounding DNA are the result of lateral gene transfer. Thus this *parB* homologue has most likely been acquired from a bacterial plasmid. The C-terminal region of ORF1-349 showed similarity to COG 5483, for which no function has yet been established. Surprisingly, a homologue of ParA/SopA appears not to be encoded by pSOG1. In bacteria, plasmid partitioning during cell division proceeds via the so-called segrosome, a protein complex at least consisting of ParA and ParB, which are encoded by the plasmid-borne *par* locus (Gerdes *et al.*, 2000; reviewed by Hayes & Barilla, 2006). ParA is a membrane-associated ATPase that forms a complex with the DNA-binding ParB (Bignell & Thomas, 2001); the binding site of ParB is the *cis*-acting partition site *parS*. The archaeal plasmid pNOB8, which is stably maintained at a low copy number in its natural host ($\sim$5 copies per chromosome), contains one ParA and two ParB homologues, as well as a putative *parS* element. These elements are missing in the unstable, high-copy-number variant pNOB8-33, formed after conjugative transfer in the foreign host *S. solfataricus* P1 (She *et al.*, 1998). Previously, it was reported that Par-like components are also absent in the other *Sulfolobus* CPs, which led to the conclusion that they may lack copy number control and partitioning (Greve *et al.*, 2004). Similarly, no homologues of *parA* or *parS* appear to be present on the genomes of the pSOG plasmids. Probably maintenance of pSOG1 does not require a partitioning system, most likely due to its high copy number (40–50 copies per chromosome) in *S. solfataricus* P1. In the case of pSOG2, however, the stable low copy number does suggest some partitioning system that is different from the Par system. One such alternative has been suggested to be the so-called clustered regularly interspaced short palindrome repeats (CRISPRs, previously referred to as SRSR). These repeats are present in many prokaryotic genomes (Jansen *et al.*, 2002), and also in pNOB and pKEF9 (Greve *et al.*, 2004; Peng *et al.*, 2003). An overexpression study in *Haloferax* initially suggested involvement in replicon partitioning (Mojica *et al.*, 1995). However, recent comparative analyses suggest a role of the repeats in a host-defence mechanism against extrachromosomal elements (viruses and plasmids) (Bolotin *et al.*, 2005; Mojica *et al.*, 2005). No CRISPRs were found in the pSOG plasmids. Hence, the molecular basis for partitioning of low-copy-number archaeal CPs, including pSOG2, remains to be identified.

Plasmid replication

The pSOG plasmids contain an operon of nine short genes (ORF113 to ORF96) presumably involved in plasmid replication. A similar operon is present in each *Sulfolobus* CP and located in conserved region C of their genome (Greve *et al.*, 2004). Five of the genes of the pSOG operon, including the first and the last two, occur in the same order in all *Sulfolobus* CPs (Fig. 4). For two of the putative proteins, searches in databases provide indirect evidence for their role in DNA replication.

Sequence similarity indicates that ORF62 belongs to the CopG family, a copy number control protein used by numerous bacterial plasmids (del Solar *et al.*, 2002). In these bacterial plasmids, the *copG* gene is located upstream of a gene encoding a replication initiator protein and the two genes are expressed from a common promoter. The CopG protein binds to this promoter and represses the expression of both proteins, thus controlling the replication of the plasmid (del Solar *et al.*, 2002). A similar organization exists in the *Sulfolobus* cryptic plasmid pRN1, where the *copG* gene precedes the gene for a RepA homologue (Keeling *et al.*, 1998). It was shown that pRN1 CopG binds to a

double-stranded DNA inverted repeat located within the *cop-rep* promoter and thus could downregulate the expression the RepA protein (Lipps *et al.*, 2001b). Such a set of inverted repeats was also identified in the promoter region of the 'replication' operon of *Sulfolobus* CPs (Greve *et al.*, 2004). In pSOG plasmids as well, a set of 8 bp inverted repeats separated by only 1 bp is found immediately downstream of the TATA box resembling the pRN1 CopG binding site and also similar to the palindromic binding site of CopG of the bacterial plasmid pMV158 (Gomis-Ruth *et al.*, 1998) (see Fig. 4 of Greve *et al.*, 2004).

A function as replication initiator protein (RepA) was proposed for one of the most conserved ORFs of the 'replication' operon (Greve *et al.*, 2004). Indeed ORF99 of pING1 shows weak but significant similarity to a putative chromosomal replication initiator of *Haemophilus ducreyi* (236 aa) and to RepA (346 aa) of plasmid pRUM from *Enterococcus faecium*. These similarities can be detected only when using the iterative PSI-BLAST tool. However, although similar in size and location in the *rep* operon, pSOG ORF106 is not homologous to the otherwise highly conserved RepA of *Sulfolobus* CPs. In fact, ORF106 of pSOG is homologous only to ORF107 of pHVE14, which is also present in the *rep* operon. Nevertheless, extensive search for a motif or conserved domain failed. Therefore no putative function could be attributed to this ORF. Since pSOG plasmids do not contain a homologue of the putative RepA-encoding gene, either it is not necessary for replication or another pSOG ORF serves that function, perhaps ORF106.

A highly conserved ORF in the 'replication' operon, ORF84, exhibits a leucine-zipper motif from position 22 to 57. This motif facilitates protein dimerization and is common to a class of DNA-binding proteins mostly found in eukaryotic transcription factors such as GCN4. A few examples of this class are known in prokaryotes, including the RepA protein of bacterial plasmids and two archaeal transcriptional regulators, GvpE of *Halobacterium salinarium* (Kruger *et al.*, 1998) and PlrA of *Sulfolobus* (discussed below).

Plasmid replication origin

There are several reasons to assume that the origin of replication (*oriV*) may be located in the region extending from ORF1-76 (and partly overlapping this ORF) to ORF87 spanning about 700 bp. First, in the Z-curve of a cumulative GC-skew analyses of pSOG plasmids (not shown) both the X and Y components show a sharp peak centred within ORF1-153 and ORF2-150 in pSOG1 and pSOG2 respectively, indicating a possible replication origin (Chen *et al.*, 2005; Zhang & Zhang, 2004). Second, that region contains a block of predicted genes, most of which are conserved in a similar gene context in other *Sulfolobus* conjugative and mobilizable plasmids. Third, this region is immediately preceded by a conspicuous AT-rich region (partly overlapping this conserved region) (Fig. 2), which may facilitate opening of the DNA strands. This region is also relatively rich in short repeated motifs that could serve as binding sites for replication factors, even though these repeats are not regularly spaced like the so-called iterons which serve as binding sites for RepA in bacterial plasmid origins (del Solar *et al.*, 1998). A putative origin of replication has been proposed for *Sulfolobus* CPs that contains a specific direct repeat 5′-TCTATACCCCC-3′ with 34–35 nt spacing in the context of a highly conserved 170 nt region (Greve *et al.*, 2004). This direct repeat with appropriate spacing is found in both pSOG1 and pSOG2 (Fig. 6), but the remainder of the sequence is not well conserved and there are substantial differences between pSOG1 and pSOG2 in both the intervening and flanking sequences. The pSOG1 sequence resembles the pKEF9 putative origin whereas the pSOG2 sequence is more similar to the pNOB8 putative origin. This may be critical for the simultaneous occurrence of both

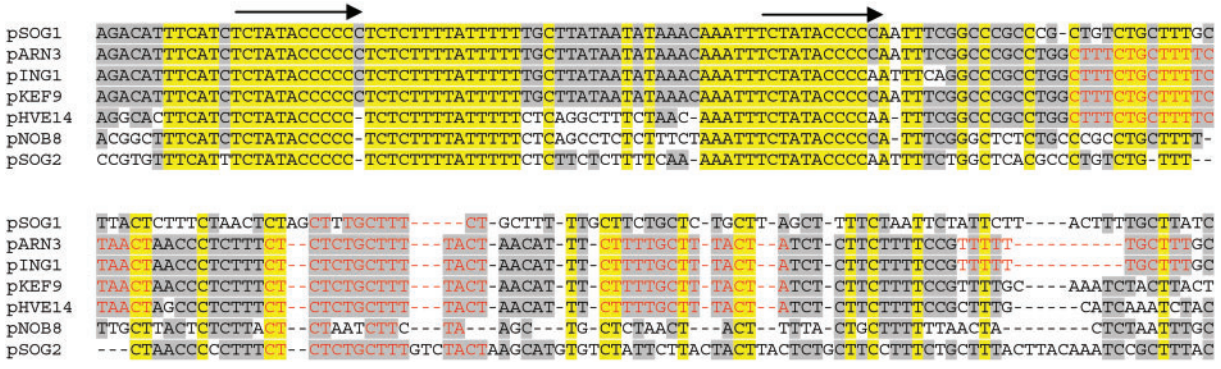

Fig. 6. Comparison of the putative *oriV* locus in pSOG1 and pSOG2 and other *Sulfolobus* CPs. Yellow background indicates blocks of nucleotides conserved in all *Sulfolobus* CPs *oriV* loci; grey background indicates those which are conserved only in a group of CPs. Black arrows indicate highly conserved 11 bp direct repeats. Red nucleotides indicate imperfect 13–18 direct repeats found in most CPs. The portion of the plasmid genome shown corresponds to the following positions: pSOG1, 13123–13307; pSOG2, 13193–13382; pARN3, 12858–13035; pING1, 13408–13585; pKEF9, 13565–13748; pNOB8, 19595–19751; pHVE14, 14997–15175.

plasmids (or of their precursors) in the original SOG2/4 strain (Fig. 1). This sequence partly overlaps with ORF1-76.

IS elements and transposases

Unlike the other *Sulfolobus* CPs, except the pARN family, pSOG plasmids do not encode any protein with homology to transposases or ORFs known to be associated with insertion sequences (Fig. 4).

Putative transcriptional regulators

Bacterial CPs have evolved systems of regulation that minimize the metabolic load on the host exerted by the maintenance of a conjugative transfer apparatus while optimizing the adaptive advantages of self-transmission. Such systems also seem to operate in *Sulfolobus* CPs, like pSOG2. Upon conjugation, pSOG2 actively replicates to a high copy number, but subsequently replication appears to be strongly down-regulated to reduce the copy number and to maintain the plasmid stably in its new host. In bacterial CPs, regulatory circuits involving specialized transcriptional regulators have been described (Zatyka & Thomas, 2002). There are as many as six ORFs that potentially play similar roles in the pSOG plasmids. The first one, ORF62 or CopG, was discussed above. ORF132 and ORF1-159 belong to a superfamily of proteins containing a winged-helix–turn–helix (wHTH) DNA-binding domain. Genomic studies have shown that this class of HTH proteins is predominant in Archaea and that its diversity is comparable to that of bacteria (for a recent review see Aravind *et al.*, 2005). Although the wHTH domain in Archaea combines with a variety of other domains including components of the replication or translation systems or in metabolic enzymes, most of the archaeal wHTH-containing proteins are predicted to be gene/operon-specific transcriptional regulators (Aravind & Koonin, 1999). This seems to be the case for ORF132 and ORF1-159, which are related to the MarR-like family (Pfam1047). Homologues of ORF132 are present in the *Sulfolobus* CP pHVE14, and several very closely related ORFs were identified in the chromosome of *S. acidocaldarius* and *S. tokodaii* (Table 1). Much weaker similarity to ORF132 was found with genes residing on pNOB8 and the *S. solfataricus* genome; no apparent homologues were found in other CPs. Interestingly, ORF1-159 has no significant similarities with other putative regulators identified in other *Sulfolobus* replicons nor with any proteins in the public databases, and its wHTH domain aligns only poorly with that of ORF132. The clear difference between the two DNA-binding proteins suggests that they have distinct functions in pSOG-related regulation. Since ORF1-159 and the closely associated ORF1-349 (ParB) were found only in pSOG1, both appear to be dispensable for *Sulfolobus* plasmids. Both ORF132 and ORF1-159 constitute a single-gene transcription unit; it is therefore difficult to infer which genes or operon they may control.

ORF78 encodes a member of the novel family of *Sulfolobus* plasmid regulatory proteins (pfam 05584) also known as PlrA (Table 1). So far representatives of this family have been found only in plasmids from the crenarchaeal genera *Sulfolobus* and *Acidianus* (Kletzin *et al.*, 1999; Peng *et al.*, 2000). This family is related to the DeoR family of bacterial transcriptional activators (Pfam 00455). It is almost identical (98 % amino acid identity) to the PlrA homologues in pARN3 and pKEF9 (Greve *et al.*, 2004), but less so (~ 50 % identical) to other PlrA proteins. One member of this new family, ORF80 of the small cryptic plasmid pRN1 from '*S. islandicus*', has been characterized (Lipps *et al.*, 2001a). It has been shown experimentally that this basic protein binds in a highly specific manner to double-stranded DNA sequences upstream of ORF80. These sequences are conserved in the region upstream of other family members including ORF78 of pSOG1. ORF80 binds DNA as a dimer. Sequence analysis suggested that this dimerization is mediated by a leucine-zipper motif, the location of which is inverted with respect to the basic domain of the protein as compared to all other known leucine-zipper proteins. ORF80 has thus been proposed to be the first representative of a novel class of leucine-zipper proteins (Lipps *et al.*, 2001a). Since the binding site of ORF80 partly overlaps with the putative archaeal TATA box, it was suggested that ORF80 represses its own transcription in an autoregulatory manner. It was suggested that ORF80 could form a complex with the replication initiation machinery (Lipps *et al.*, 2001a). Moreover, it has been proposed that the region upstream of ORF80 contains the double-stranded origin of replication in pRN1, and that ORF80 could be involved in the regulation of plasmid copy number (Kletzin *et al.*, 1999; Peng *et al.*, 2000). Experimental evidence supporting these hypotheses is still lacking. All other plasmids of *Sulfolobus*, with the exception of pORA1 and pSOG2, contain PlrA homologues (Greve *et al.*, 2004). This indicates an important but not essential role for PlrA for *Sulfolobus* plasmid function.

Interestingly, ORF93a and ORF175a, which form a putative operon with ORF211 (in the order 93a–211–175a), are also wHTH-containing proteins. ORF175a belongs to the TrmB family (Pfam 01978), of which two members have recently been characterized: TrmB is a sugar-specific transcriptional regulator of the operon encoding the trehalose/maltose ABC transporter in the hyperthermophilic euryarchaea *Thermococcus litoralis* and *Pyrococcus furiosus* (Lee *et al.*, 2003). ORF93a belongs to a small family of predicted transcriptional regulator proteins of euryarchaeotes. ORF175a is not conserved in other described *Sulfolobus* CPs but is present in the sequence deposed in GenBank (NC_005969) for plasmid pTC of '*Sulfolobus. tengchongensis*' and in the integrated plasmid-related element SA3 found in the chromosome of *S. acidocaldarius* (Chen *et al.*, 2005). ORF93a homologues are found only in pARN3 and pHVE14. However in the other CPs (pKEF9, pING1 and pNOB8) an ORF of about the same size (encoding 92–95 aa) is found upstream of the ORF211 homologues (conserved in all the CPs) forming a putative operon. All of these ORFs are predicted to encode archaeal transcriptional regulators of the wHTH clan related either to the MarR, LysR or to the ArsR families. Thus, all

these small regulators seem to be interchangeable as long as they ensure the same function in the same genomic context in different plasmids.

The integrase of pSOG plasmids and its integration site

Like other CPs of *Sulfolobus*, both pSOG plasmids contain a homologue (ORF421) of a new family of integrase proteins, the pNOB8-type integrase, originally identified by comparing plasmid-like sequences in the *S. tokodaii* genome (Kawarabayasi *et al.*, 2001) with the pNOB8 CP (She *et al.*, 2002). These integrases are assumed to facilitate reversible integration of the CPs into the host chromosome by site-specific recombination between a plasmid attachment site *attP* and the corresponding chromosomal *attB* site. The pNOB8-type integrase belongs to the superfamily of tyrosine recombinases, which play several crucial roles in prokaryotes and eukaryotes (for a review see Van Duyne, 2002). One striking feature of this family is the lack of global homology among its more than 150 members. Nevertheless, a conserved signature is found in the C-terminal part of all the proteins. All members of the family harbour two short regions of similarity, box I and box II, sharing four nearly invariant amino acids residues, R…HxxR…Y, directly involved in catalysis of the DNA strand cleavage and exchange (Esposito & Scocca, 1997; Nünes-Duby *et al.*, 1998). Sequence alignments revealed that both the motifs found for the integrase of the *Sulfolobus* virus SSV1, the SSV-type, R…Kxxx…Y, and for the pNOB8-type integrase, R…Yxxx…Y differ from the consensus and may represent three major classes of integrase (She *et al.*, 2004).

The overall sequence of pSOG integrase is similar to that of other *Sulfolobus* CPs and to the related integrase genes found in the *Sulfolobus* and *Aeropyrum* chromosomes (an alignment of the *Sulfolobus* CPs integrases with representative members of the tyrosine recombinases is available as supplementary data with the online version of this paper). Among these, however, pSOG integrase is more closely related to the integrases of pKEF9, pARN3 and pHVE14, with which it shares up to 63 % identity. These integrase sequences clearly form a homogeneous distinct group. As expected from the general features of tyrosine recombinases, the differences from the other aligned sequences are more pronouced in their N-terminal halves. Interestingly, search for conserved protein domains revealed that pSOG Int and the five other closely related integrases harbour a typical HTH-XRE domain (smart: SM00530, pfam: PF01381) in their N-termini (amino acid positions 17–69 in pSOG Int). This protein domain is found in a large family of DNA-binding proteins that include a bacterial plasmid copy control protein, bacterial methylases, and various bacteriophage transcription control proteins like the Cro and cI repressors of bacteriophage λ. In Archaea, this motif is also well represented, with 106 entries in the Smart database. Most of them are small proteins (less than than 200 aa) with no other conserved domain and are predicted to be transcriptional regulators. Several of the archaeal HTH-XRE-containing proteins possess additional enzymic or protein–protein interaction domains. In a few cases, a HTH-XRE domain is fused to a metabolic enzyme (e.g. purine phosphoribosyltransferase of *Pyrococcus abyssi*, PAB2035). These proteins might combine the catalytic function with that of transcription regulation of the biosynthetic genes in response to the respective metabolite in the environment. Another type of association is found in archaeal inteins (e.g. those of the replication factor C from *Methanococcus jannaschii*), where the HTH-XRE domain is sandwiched between the N-terminal part of the intein module and the inserted homing endonuclease domain (Gogarten *et al.*, 2002). This association with the endonuclease domain suggests that the HTH might play a role in the recognition of target sequences by the endonuclease in the process of homing.

By analogy with the previous examples, we envisage two general roles for the HTH domain in the pSOG integrase. First, the HTH domain could contribute to the integrase DNA-binding at the *att*P site. In the prototype integrase, the λ integrase, the 356 aa sequence can be split into two domains by limited proteolysis. The N-terminal domain includes residues 1–64 and is responsible for binding the so-called arm-type sites of *attP* [adjacent direct repeats sites that flank the core region where crossing-over occurs (Groth & Calos, 2004)] while the C-terminal domain binds the lower-affinity core-type sites and contains the catalytic site (Groth & Calos, 2004). That the HTH domain of pSOG integrase could serve in binding of arm-type sites seems unlikely since such a domain is absent in the closely related pNOB8 integrase which has recently been proved active in *S. solfataricus* (She *et al.*, 2004), indicating that this domain is not essential for the activity of the protein. Moreover, none of the identified prokaryotic or eukaryotic integrases possesses a HTH domain in its N-terminus. We therefore infer that the HTH-XRE motif is somehow involved in transcriptional regulation of the integration/excision of pSOG plasmid.

The putative integration site *attP* used by the pSOG plasmids corresponds to a 43 bp invariant sequence which is identical to the 3′ end of two glutamyl-tRNA genes in the *S. solfataricus* P2 genome. In virus SSV1, the only well-studied archaeal integration system, a conserved 44 bp sequence, identical to the 3′ half of an arginyl-tRNA gene, was found in the genome of the host *Sulfolobus shibatae* flanking the provirus as direct repeats (Muskhelishvili *et al.*, 1993). Recent studies showed that the SSV1 integrase cleaves both DNA strands at the *att* sites and that the cleavage positions are localized on each side of the anti-codon loop of the tRNA where SSV1 integration takes place (Serre *et al.*, 2002). This situation occurs quite frequently in the prokaryotic integrases, where some subfamilies recognize the flanking symmetry of the anti-codon stem–loop structure and use exclusively this tRNA sublocation as integration site (Williams, 2002). Alignment of the *Sulfolobus* CP *attP* sequences with those of the corresponding

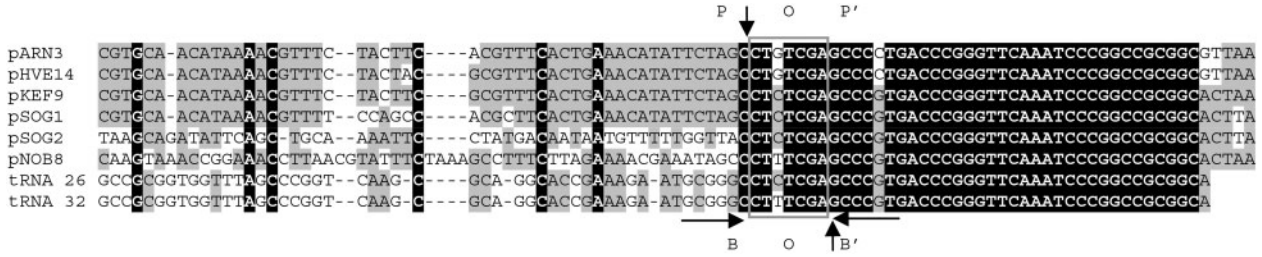

Fig. 7. Alignment of the *Sulfolobus* CP integration sites. Conserved sequence positions are indicated on a black (completely conserved) or grey (partly conserved) background. Sequences in the *attP* region of each CP are aligned with their cognate tRNA sequence from the *S. solfataricus* genome (no site found for pING1). The boxed sequence corresponds to the tRNA anticodon loop and the flanking vertical arrows indicate the putative integrase cleavage positions. The core site symmetrical elements of *attP* (P, P′) and *attB* (B, B′) are indicated following the conventions used by Campbell (1992) for bacterial integration sites. The portion of the plasmid genome shown corresponds to the following positions: pSOG1, 21418–21516; pSOG2, 22791–22889; pARN3, 23442–23540; pKEF9, 25448–25546; pNOB8, 31203–31109; pHVE14, 27790–27867. The portions of sequences shown are the reverse complement of that deposited in GenBank.

tRNA genes in *S. solfatatricus* (Fig. 7) strongly suggests that all the *Sulfolobus* CP integrases also use the anti-codon stem–loop sublocation as integration site and therefore belong to the same class of integrases as SSV1 Int. Interestingly, several *Sulfolobus* CP integrases, including pSOG and pNOB8, apparently share the same putative integration sites in *Sulfolobus*, the two $tRNA^{Glu}$ genes. However, pSOG integrase may use preferentially the $tRNA^{Glu}$ with a CTC anti-codon (Ssot26) that shows a perfect match with the pSOG *attP*, while the second (Ssot32), with TTC as anti-codon, has one mismatch in the anticodon. This situation is the exact opposite of that in pNOB8 integrase (She *et al.*, 2004). Surprisingly, no *attP* site was detected in pING1 plasmids, which therefore can no longer integrate into the host chromosome.

Comparison of the sites of recombination in the pSOG1 and pSOG2 genomes

A putative recombination motif was previously described to explain the variation of the pING family of *Sulfolobus* CPs (Stedman *et al.*, 2000). This motif, 5′-TAAACTGGGGAG-TTA-3′, was also found in regions of sequence divergence in the *Sulfolobus* CPs pHVE14, pKEF9, pARN3, pARN4 and pNOB8 (Greve *et al.*, 2004). Strikingly, this motif is found flanking all but one of the locations in the pSOG1 and pSOG2 genomes in which they differ (Fig. 2, Fig. 4). Two of these sequences flank the conserved block of ORFs predicted to have conjugative functions (Fig. 4, section A) and presumably allowed the recombination event producing pSOG1 and pSOG2. Other recombination motifs flank the region near the putative replication origin and the identical sequences in pSOG1 and pSOG2. The only major gene insertion in pSOG1 that does not contain this flanking motif is between pSOG1 ORFs 1-421 and 1-288. This is the region that encompasses the PlrA and ParB homologues in pSOG1, leading us to speculate that this region was deleted from pSOG2 rather than inserted in pSOG1. There are no other sequences surrounding this region that indicate recombination, other than that of the flanking integrase gene ORF421.

Plasmid copy number and stability

The development of genetic tools for hyperthermophilic Archaea in general and *Sulfolobus* in particular has been hampered by the relative lack of stable plasmids with controlled copy number. The fuselloviruses of *Sulfolobus* replicate as double-stranded circular DNA and have been used as self-spreading plasmids (Jonuscheit *et al.*, 2003; Schleper *et al.*, 1995; Stedman *et al.*, 1999). However, with larger insertions these plasmids are not very stable (Jonuscheit *et al.*, 2003). Their copy number control is also not well understood. The large *Sulfolobus* CP pNOB8 was used as a vector for the first successful transformation of *Sulfolobus* (Elferink *et al.*, 1996) but is also not stable.

This study was initiated in order to determine the genetic basis for stability and copy number control of the pSOG plasmids. Plasmid pSOG2 is very attractive as a potential vector as it has a stable low copy number in *S. solfataricus* P1 and can be transferred from cell to cell by conjugation. Unfortunately the molecular basis of this control is not clear. Perhaps the presence of the PlrA protein in pSOG1 causes a higher copy number or the origin of replication of pSOG1 is more active. Stability may be directly related to plasmid copy number, as strains containing CPs grow much more slowly than those that do not (Prangishvili *et al.*, 1998; Schleper *et al.*, 1995).

Plasmid compatibility and use as vectors

Plasmid compatibility has not been well studied in *Sulfolobus*. Compatible replicons are critical for sophisticated genetic experiments. The integration sites of different SSV viruses are different, indicating that they may be compatible with each other, but this has yet to be demonstrated (Wiedenheft *et al.*, 2004). It is possible to co-infect certain '*S. islandicus*' strains with both the SIRV virus and SSV1

(Prangishvili *et al.*, 1999). Small plasmids can occur in the presence of either larger plasmids or virus genomes, as has been shown for SSV2 and the virus/plasmid hybrid pSSVx (Arnold *et al.*, 1999). The non-conjugative plasmids pRN1 and pRN2 were both found in the same strain of '*S. islandicus*', REN1H1 (Zillig *et al.*, 1994), but contain no selectable markers. The conjugation proteins of pSOG1 and its putative replication origin are clearly related to the pKEF family of *Sulfolobus* CPs, whereas the conjugation proteins and the origin of pSOG2 are clearly related to counterparts of the pARN family of *Sulfolobus* CPs. These two plasmids or precursors thereof are found in the SOG2/4 strain ('pSOG1' and 'pSOG2/4') (Fig. 1.). These two plasmids are the first example of two different families of *Sulfolobus* CPs to be found in the same *Sulfolobus* strain. Compatibility between different families of *Sulfolobus* CPs was previously demonstrated in laboratory conjugation experiments but had not previously been shown to exist in naturally occurring strains (Prangishvili *et al.*, 1998). It remains to be determined if these CPs are also compatible with *Sulfolobus* viruses and small plasmids. In any case pSOG2 should be a useful addition to the *Sulfolobus* genetics tool-kit as a low-copy-number stable CP. pSOG1 may be useful as a 'Trojan horse' for the introduction of manipulated genes into host chromosomes by homologous recombination or transient expression.

ACKNOWLEDGEMENTS

G. E. was supported by EU grant ERBIO4-CT96-0270; K. S. was supported by a Marie Curie Fellowship from the European Commission, a NSF-NATO Fellowship and Portland State University. Thanks to members of the Zillig lab for Southern blots on '*S. islandicus*' strains.

REFERENCES

Albers, S.-V., Jonuscheit, M., Dinkelaker, S., Urich, T., Kletzin, A., Tampe, R., Driessen, A. J. M. & Schleper, C. (2006). Production of recombinant and tagged proteins in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Appl Environ Microbiol* **72**, 102–111.

Aravalli, R. N. & Garrett, R. A. (1997). Shuttle vectors for hyperthermophilic archaea. *Extremophiles* **1**, 183–191.

Aravind, L. & Koonin, E. V. (1999). DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res* **27**, 4658–4670.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* **29**, 231–262.

Arnold, H. P., She, Q., Phan, H., Stedman, K., Prangishvili, D., Holz, I., Kristjansson, J. K., Garrett, R. & Zillig, W. (1999). The genetic element pSSVx of the extremely thermophilic crenarchaeon *Sulfolobus* is a hybrid between a plasmid and a virus. *Mol Microbiol* **34**, 217–226.

Bartolucci, S., Rossi, M. & Cannio, R. (2003). Characterization and functional complementation of a nonlethal deletion in the chromosome of a β-glycosidase mutant of *Sulfolobus solfataricus*. *J Bacteriol* **185**, 3948–3957.

Bell, S. D. & Jackson, S. P. (2000). The role of transcription factor B in transcription initiation and promoter clearance in the archaeon *Sulfolobus acidocaldarius*. *J Biol Chem* **275**, 12934–12940.

Benes, V., Hostomsky, Z., Arnold, L. & Paces, V. (1993). M13 and pUC vectors with new unique restriction sites for cloning. *Gene* **130**, 151–152.

Bignell, C. & Thomas, C. M. (2001). The bacterial ParA-ParB partitioning proteins. *J Biotechnol* **91**, 1–34.

Birboim, H. C. & Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res* **7**, 1513–1523.

Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561.

Brugger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y. & Garrett, R. A. (2002). Mobile elements in archaeal genomes. *FEMS Microbiol Lett* **206**, 131–141.

Campbell, A. (1992). Chromosomal insertion sites for phages and plasmids. *J Bacteriol* **174**, 7495–7499.

Cannio, R., Contursi, P., Rossi, M. & Bartolucci, S. (1998). An autonomously replicating transforming vector for *Sulfolobus solfataricus*. *J Bacteriol* **180**, 3237–3240.

Chen, L., Brugger, K., Skovgaard, M. & 8 other authors (2005). The Genome of *Sulfolobus acidocaldarius*, a model organism of the Crenarchaeota. *J Bacteriol* **187**, 4992–4999.

Condo, I., Ciammaruconi, A., Benelli, D., Ruggero, D. & Londei, P. (1999). Cis-acting signals controlling translational initiation in the thermophilic archaeon *Sulfolobus solfataricus*. *Mol Microbiol* **34**, 377–384.

Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190.

Dalgaard, J. Z. & Garrett, R. A. (1993). Archaeal hyperthermophile genes. In *The Biochemistry of Archaea (Archaebacteria)*, pp. 535–563. Edited by M. Kates, D. J. Kusher & A. T. Matheson. Amsterdam: Elsevier.

del Solar, G., Giraldo, R., Ruiz-Echevarria, M.-J., Espinosa, M. & Diaz-Orejas, R. (1998). Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev* **62**, 434–464.

del Solar, G., Hernandez-Arriaga, A. M., Gomis-Ruth, F. X., Coll, M. & Espinosa, M. (2002). A genetically economical family of plasmid-encoded transcriptional repressors involved in control of plasmid copy number. *J Bacteriol* **184**, 4943–4951.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797.

Elferink, M. G., Schleper, C. & Zillig, W. (1996). Transformation of the extremely thermoacidophilic archaeon *Sulfolobus solfataricus* via a self-spreading vector. *FEMS Microbiol Lett* **137**, 31–35.

Elie, C., Baucher, M. F., Fondrat, C. & Forterre, P. (1997). A protein related to eucaryal and bacterial DNA-motor proteins in the hyperthermophilic archaeon *Sulfolobus acidocaldarius*. *J Mol Evol* **45**, 107–114.

Esposito, D. & Scocca, J. (1997). The integrase family of tyrosine recombinases: evolution of a conserved active site domain. *Nucleic Acids Res* **25**, 3605–3614.

Francia, M. V., Varsaki, A., Garcillan-Barcia, M. P., Latorre, A., Drainas, C. & de la Cruz, F. (2004). A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiol Rev* **28**, 79–100.

Garcia-Vallve, S., Guzman, E., Montero, M. A. & Romeu, A. (2003). HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res* **31**, 187–189.

Gerdes, K., Moller-Jensen, J. & Bugge Jensen, R. (2000). Plasmid and chromosome partitioning: surprises from phylogeny. *Mol Microbiol* **37**, 455–466.

Gogarten, J. P., Senejani, A. G., Zhaxybayeva, O., Olendzenski, L. & Hilario, E. (2002). Inteins: structure, function, and evolution. *Annu Rev Microbiol* **56**, 263–287.

Gomis-Ruth, F. X., Sola, M., Acebo, P. & 7 other authors (1998). The structure of plasmid-encoded transcriptional repressor CopG unliganded and bound to its operator. *EMBO J* **17**, 7404–7415.

Grahn, A. M., Haase, J., Bamford, D. H. & Lanka, E. (2000). Components of the RP4 conjugative transfer apparatus form an envelope structure bridging inner and outer membranes of donor cells: implications for related macromolecule transport systems. *J Bacteriol* **182**, 1564–1574.

Greve, B., Jensen, G. B., Brügger, K., Zillig, W. & Garrett, R. (2004). Genomic comparison of archaeal plasmids from *Sulfolobus*. *Archaea* **1**, 231–239.

Grohmann, E., Muth, G. & Espinosa, M. (2003). Conjugative plasmid transfer in gram-positive bacteria. *Microbiol Mol Biol Rev* **67**, 277–301.

Groth, A. C. & Calos, M. P. (2004). Phage integrases: biology and applications. *J Mol Biol* **335**, 667–678.

Hanai, R., Liu, R., Benedetti, P., Caron, P. R., Lynch, A. S. & Wang, J. C. (1996). Molecular dissection of a protein SopB essential for *Escherichia coli* F plasmid partition. *J Biol Chem* **271**, 17469–17475.

Hayes, F. & Barilla, D. (2006). The bacterial segrosome: a dynamic nucleoprotein machine for DNA trafficking and segregation. *Nat Rev Micro* **4**, 133–143.

Jansen, R., Embden, J. D. A. V., Gaastra, W. & Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565–1575.

Jonuscheit, M., Martusewitsch, E., Stedman, K. M. & Schleper, C. (2003). A reporter gene system for the hyperthermophilic archaeon *Sulfolobus solfataricus* based on a selectable and integrative shuttle vector. *Mol Microbiol* **48**, 1241–1252.

Kawarabayasi, Y., Hino, Y., Horikawa, H. & 27 other authors (2001). Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain 7. *DNA Res* **8**, 123–140.

Keeling, P. J., Klenk, H. P., Singh, R. K., Schenk, M. E., Sensen, C. W., Zillig, W. & Doolittle, W. F. (1998). *Sulfolobus islandicus* plasmids pRN1 and pRN2 share distant but common evolutionary ancestry. *Extremophiles* **2**, 391–393.

Kletzin, A., Lieke, A., Urich, T., Charlebois, R. L. & Sensen, C. W. (1999). Molecular analysis of pDL10 from *Acidianus ambivalens* reveals a family of related plasmids from extremely thermophilic and acidophilic archaea. *Genetics* **152**, 1307–1314.

Kruger, K., Hermann, T., Armbruster, V. & Pfeifer, F. (1998). The transcriptional activator GvpE for the halobacterial gas vesicle genes resembles a basic region leucine-zipper regulatory protein. *J Mol Biol* **279**, 761–771.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150–163.

Lanka, E. & Wilkins, B. M. (1995). DNA processing reactions in bacterial conjugation. *Annu Rev Biochem* **64**, 141–169.

Lee, S. J., Engelmann, A., Horlacher, R., Qu, Q., Vierke, G., Hebbeln, C., Thomm, M. & Boos, W. (2003). TrmB, a sugar-specific transcriptional regulator of the trehalose/maltose ABC transporter from the hyperthermophilic archaeon *Thermococcus litoralis*. *J Biol Chem* **278**, 983–990.

Lipps, G., Ibanez, P., Stroessenreuther, T., Hekimian, K. & Krauss, G. (2001a). The protein ORF80 from the acidophilic and thermophilic archaeon *Sulfolobus islandicus* binds highly site-specifically to double-stranded DNA and represents a novel type of basic leucine zipper protein. *Nucleic Acids Res* **29**, 4973–4982.

Lipps, G., Stegert, M. & Krauss, G. (2001b). Thermostable and site-specific DNA binding of the gene product ORF56 from the *Sulfolobus islandicus* plasmid pRN1, a putative archael plasmid copy control protein. *Nucleic Acids Res* **29**, 904–913.

Llosa, M. & de la Cruz, F. (2005). Bacterial conjugation: a potential tool for genomic engineering. *Res Microbiol* **156**, 1–6.

Llosa, M., Gomis-Ruth, F. X., Coll, M. & de la Cruz Fd, F. (2002). Bacterial conjugation: a two-step mechanism for DNA transport. *Mol Microbiol* **45**, 1–8.

Llosa, M., Zunzunegui, S. & de la Cruz, F. (2003). Conjugative coupling proteins interact with cognate and heterologous VirB10-like proteins while exhibiting specificity for cognate relaxosomes. *Proc Natl Acad Sci U S A* **100**, 10465–10470.

Martusewitsch, E., Sensen, C. W. & Schleper, C. (2000). High spontaneous mutation rate in the hyperthermophilic archaeon *Sulfolobus solfataricus* is mediated by transposable elements. *J Bacteriol* **182**, 2574–2581.

Mojica, F. J., Ferrer, C., Juez, G. & Rodriguez-Valera, F. (1995). Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. *Mol Microbiol* **17**, 85–93.

Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. C. & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174–182.

Muskhelishvili, G., Palm, P. & Zillig, W. (1993). SSV1-encoded site-specific recombination system in *Sulfolobus shibatae*. *Mol Gen Genet* **237**, 334–342.

Nei, M. & Kumar, K. (2000). *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Nünes-Duby, S. E., Kwon, H. J., Tirumalai, R. S., Ellenberger, T. & Landy, A. (1998). Similarities and differences among 105 members of the *Int* family of site-specific recombinases. *Nucleic Acids Res* **26**, 391–406.

Palm, P., Schleper, C., Grampp, B., Yeats, S., McWilliam, P., Reiter, W. D. & Zillig, W. (1991). Complete nucleotide sequence of the virus SSV1 of the archaebacterium *Sulfolobus shibatae*. *Virology* **185**, 242–250.

Pansegrau, W., Lanka, E., Barth, P. T. & 7 other authors (1994). Complete nucleotide sequence of Birmingham IncPα plasmids. Compilation and comparative analysis. *J Mol Biol* **239**, 623–663.

Peng, X., Holz, I., Zillig, W., Garrett, R. A. & She, Q. (2000). Evolution of the family of pRN plasmids and their integrase-mediated insertion into the chromosome of the crenarchaeon *Sulfolobus solfataricus*. *J Mol Biol* **303**, 449–454.

Peng, X., Brugger, K., Shen, B., Chen, L., She, Q. & Garrett, R. A. (2003). Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes. *J Bacteriol* **185**, 2410–2417.

Prangishvili, D., Albers, S. V., Holz, I. & 8 other authors (1998). Conjugation in Archaea: frequent occurrence of conjugative plasmids in *Sulfolobus*. *Plasmid* **40**, 190–202.

Prangishvili, D., Arnold, H. P., Gotz, D., Ziese, U., Holz, I., Kristjansson, J. K. & Zillig, W. (1999). A novel virus family, the Rudiviridae: structure, virus-host interactions and genome variability of the sulfolobus viruses SIRV1 and SIRV2. *Genetics* **152**, 1387–1396.

Prangishvili, D., Stedman, K. & Zillig, W. (2001). Viruses of the extremely thermophilic archaeon *Sulfolobus*. *Trends Microbiol* **9**, 39–43.

Reilly, M. S. & Grogan, D. W. (2001). Characterization of intragenic recombination in a hyperthermophilic archaeon via conjugational DNA exchange. *J Bacteriol* **183**, 2943–2946.

Reiter, W. D., Palm, P. & Ziilig, W. (1988). Analysis of transcription in the archaebacterium *Sulfolobus* indicates that archaebacterial promoters are homologous to eukaryotic pol II promoters. *Nucleic Acids Res* **16**, 1–19.

Rice, G., Stedman, K., Snyder, J., Wiedenheft, B., Willits, D., Brumfield, S., McDermott, T. & Young, M. J. (2001). Viruses from extreme thermal environments. *Proc Natl Acad Sci U S A* **98**, 13341–13345.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Schleper, C., Kubo, K. & Zillig, W. (1992). The particle SSV1 from the extremely thermophilic archaeon *Sulfolobus* is a virus: demonstration of infectivity and of transfection with viral DNA. *Proc Natl Acad Sci U S A* **89**, 7645–7649.

Schleper, C., Holz, I., Janekovic, D., Murphy, J. & Zillig, W. (1995a). A multicopy plasmid of the extremely thermophilic archaeon *Sulfolobus* effects its transfer to recipients by mating. *J Bacteriol* **177**, 4417–4426.

Schroder, G. & Lanka, E. (2003). TraG-like proteins of Type IV secretion systems: functional dissection of the multiple activities of TraG (RP4) and TrwB (R388). *J Bacteriol* **185**, 4371–4381.

Schroder, G., Krause, S., Zechner, E. L., Traxler, B., Yeo, H.-J., Lurz, R., Waksman, G. & Lanka, E. (2002). TraG-like proteins of DNA transfer systems and of the *Helicobacter pylori* Type IV secretion system: inner membrane gate for exported substrates? *J Bacteriol* **184**, 2767–2779.

Serre, M. C., Letzelter, C., Garel, J. R. & Duguet, M. (2002). Cleavage properties of an archaeal site-specific recombinase, the SSV1 integrase. *J Biol Chem* **277**, 16758–16767.

She, Q., Phan, H., Garrett, R. A., Albers, S. V., Stedman, K. M. & Zillig, W. (1998). Genetic profile of pNOB8 from *Sulfolobus*: the first conjugative plasmid from an archaeon. *Extremophiles* **8**, 417–425.

She, Q., Singh, R. K., Confalonieri, F. & 28 other authors (2001). The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc Natl Acad Sci U S A* **98**, 7835–7840.

She, Q., Brugger, K. & Chen, L. (2002). Archaeal integrative genetic elements and their impact on genome evolution. *Res Microbiol* **153**, 325–332.

She, Q., Shen, B. & Chen, L. (2004). Archaeal integrases and mechanisms of gene capture. *Biochem Soc Trans* **32**, 222–226.

Soppa, J. (1999). Normalized nucleotide frequencies allow the definition of archaeal promoter elements for different archaeal groups and reveal base-specific TFB contacts upstream of the TATA box. *Mol Microbiol* **31**, 1589–1592.

Stedman, K. M., Schleper, C., Rumpf, E. & Zillig, W. (1999). Genetic requirements for the function of the archaeal virus SSV1 in *Sulfolobus solfataricus*: construction and testing of viral shuttle vectors. *Genetics* **152**, 1397–1405.

Stedman, K. M., She, Q., Phan, H., Holz, I., Singh, H., Prangishvili, D., Garrett, R. & Zillig, W. (2000). pING family of conjugative plasmids from the extremely thermophilic archaeon *Sulfolobus islandicus*: insights into recombination and conjugation in *Crenarchaeota*. *J Bacteriol* **182**, 7014–7020.

Thompson, W., Rouchka, E. C. & Lawrence, C. E. (2003). Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* **31**, 3580–3585.

Tolstrup, N., Sensen, C. W., Garrett, R. A. & Clausen, I. G. (2000). Two different and highly organized mechanisms of translation initiation in the archaeon *Sulfolobus solfataricus*. *Extremophiles* **4**, 175–179.

Torarinsson, E., Klenk, H.-P. & Garrett, R. A. (2005). Divergent transcriptional and translational signals in Archaea. *Environ Microbiol* **7**, 47–54.

Van Duyne, G. D. (2002). A structural view of tyrosine recombinase site-specific recombination. In *Mobile DNA II*, pp. 93–117. Edited by N. L. Craig, R. Craigie, M. Gellert & A. Lambowitz. Washington, DC: American Society for Microbiology.

Wiedenheft, B., Stedman, K., Roberto, F., Willits, D., Gleske, A.-K., Zoeller, L., Snyder, J., Douglas, T. & Young, M. (2004). Comparative genomic analysis of hyperthermophilic archaeal fuselloviridae viruses. *J Virol* **78**, 1954–1961.

Williams, K. P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res* **30**, 866–875.

Worthington, P., Hoang, V., Perez-Pomares, F. & Blum, P. (2003). Targeted disruption of the α-amylase gene in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *J Bacteriol* **185**, 482–488.

Xiang, X., Dong, X. & Huang, L. (2003). *Sulfolobus tengchongensis* sp. nov., a novel thermoacidophilic archaeon isolated from a hot spring in Tengchong, China. *Extremophiles* **7**, 493–498.

Zatyka, M. & Thomas, C. M. (2002). Control of genes for conjugative transfer of plasmids and other mobile elements. *FEMS Microbiol Rev* **21**, 291–319.

Zhang, R. & Zhang, C. T. (2004). Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea* **1**, 335–346.

Zillig, W., Stetter, K. O., Wunderl, S., Schulz, W., Priess, H. & Scholz, I. (1980). The *Sulfolobus* -"*Caldariella*" group: taxonomy on the basis of the structure of DNA-dependent RNA polymerases. *Arch Microbiol* **125**, 259–269.

Zillig, W., Kletzin, A., Schleper, C., Holz, I., Janecovik, D., Hain, J., Lanzendorfer, M. & Kristjansson, J. (1994). Screening for *Sulfolobales*, their plasmids and viruses in Icelandic solfataras. *Syst Appl Microbiol* **16**, 609–628.

Zillig, W., Arnold, H. P., Holz, I. & 7 other authors (1998). Genetic elements in the extremely thermophilic archaeon *Sulfolobus*. *Extremophiles* **2**, 131–140.